Journal of the American Medical Informatics Association, 25(3), 2018, 345–352
doi: 10.1093/jamia/ocx137
Advance Access Publication Date: 1 December 2017
Research and Applications

Research and Applications

PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data

Jing Huang,[1,*] Rui Duan,[1,*] Rebecca A Hubbard,[1] Yonghui Wu,[2] Jason H Moore,[1] Hua Xu,[2] and Yong Chen[1]

[1]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA and [2]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA

Corresponding Author: Yong Chen, School of Medicine, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104, USA. E-mail: ychen123@upenn.edu. Phone: 215-746-8155

*The first two authors contributed equally.
Received 23 March 2017; Revised 10 October 2017; Editorial Decision 20 October 2017; Accepted 15 November 2017

ABSTRACT

Objectives: This study proposes a novel Prior knowledge guided Integrated likelihood Estimation (PIE) method to correct bias in estimations of associations due to misclassification of electronic health record (EHR)-derived binary phenotypes, and evaluates the performance of the proposed method by comparing it to 2 methods in common practice.
Methods: We conducted simulation studies and data analysis of real EHR-derived data on diabetes from Kaiser Permanente Washington to compare the estimation bias of associations using the proposed method, the method ignoring phenotyping errors, the maximum likelihood method with misspecified sensitivity and specificity, and the maximum likelihood method with correctly specified sensitivity and specificity (gold standard). The proposed method effectively leverages available information on phenotyping accuracy to construct a prior distribution for sensitivity and specificity, and incorporates this prior information through the integrated likelihood for bias reduction.
Results: Our simulation studies and real data application demonstrated that the proposed method effectively reduces the estimation bias compared to the 2 current methods. It performed almost as well as the gold standard method when the prior had highest density around true sensitivity and specificity. The analysis of EHR data from Kaiser Permanente Washington showed that the estimated associations from PIE were very close to the estimates from the gold standard method and reduced bias by 60%–100% compared to the 2 commonly used methods in current practice for EHR data.
Conclusions: This study demonstrates that the proposed method can effectively reduce estimation bias caused by imperfect phenotyping in EHR-derived data by incorporating prior information through integrated likelihood.

Key words: association study, bias reduction, electronic health record, misclassification, prior information.

INTRODUCTION

Electronic health records (EHRs) have emerged as a major source of data for clinical and health services research.[1–5] Despite their great potential, the complex and inconsistent nature of EHR data brings additional challenges for many clinical studies. One such challenge is information bias, also known as observation, classification, or measurement bias, which results from incorrect determination of outcomes, exposures, or both in EHR-derived data.[6–8] In particular, automated phenotyping algorithms, which extract patients' disease,

treatment, and response information from EHRs using both structured data (eg, International Classification of Diseases, Ninth and Tenth Revision codes) and unstructured data (eg, clinical narratives) through advanced informatics technologies, may create misclassification or measurement errors due to limited sensitivity and specificity of the algorithms.[9–11] Current practice in EHR-based studies usually requires that phenotyping algorithms achieve reasonable performance[1,5] but ignores the errors of EHR phenotyping in subsequent analysis, which could lead to biased estimations of associations and loss of power in further association studies.[12] Recently, we conducted extensive simulation studies motivated by real-world EHR data to quantify power loss due to misclassification of binary outcomes in EHR-based genetic and epidemiological association studies.[13] We explored various settings, including different levels of sensitivity and specificity, and found that estimation bias and power loss can be substantial. Even in a relatively low misclassification situation, where the positive predictive value of the algorithm is 0.90 and the sensitivity is 0.84, the power loss can be as much as 25% due to misclassification.[13] Alternatively, phenotyping can be conducted by manual chart review, but this approach is time-consuming and costly. In most situations, only a small validation study using manual chart review is affordable.

Standard likelihood-based or Bayesian methods can address this challenge by accounting for misclassification and measurement errors.[14] Specifically, information bias due to an imperfect phenotyping algorithm can be parameterized using phenotype misclassification parameters (sensitivity and specificity), and association estimates can be obtained using joint estimates of the misclassification parameters and the association parameters by a maximum likelihood (ML) or Bayesian approach.[12,15,16] However, the sample size required to successfully carry out this joint estimation is very large, making this approach impracticable.[14] Intuitively, the identified "cases" and "controls" are a mixture of both diseased and healthy individuals. Thus, joint estimation of the misclassification parameters and the association parameters is a mixture-model problem, which is notoriously difficult and requires an extremely large sample size. In practice, investigators conducting EHR-based studies have found that the maximum likelihood estimator (MLE) of the association parameters using joint estimation has large bias and variability.[17] To overcome this challenge, one approach is to fix the sensitivity and specificity at given values and maximize the likelihood with respect to the association parameters only. Such methods require the use of validation data to estimate the sensitivity and specificity or the use of previously reported misclassification rates directly.[12] Two main limitations of these methods are as follows: First, to achieve unbiasedness in the estimation of association parameters, the validation sample must be large, but in practical EHR-based studies, the validation sample size is typically relatively small (eg, 100–500). Second, the performance of a phenotyping algorithm may vary substantially when applied to a different EHR dataset from the one for which it was originally developed.[18,19] Direct use of literature-reported misclassification rates may cause incorrect specification of the parameters, which also leads to biased estimation of associations.[20]

In this paper, we propose a novel Prior knowledge guided Integrated likelihood Estimation method (PIE) to address the challenge of information bias caused by phenotyping errors without specifying fixed values for the sensitivity and specificity of phenotyping algorithms. The proposed method incorporates prior knowledge about phenotype sensitivity and specificity through integrated likelihood (IL),[21] where uncertainty in sensitivity and specificity is rigorously accounted for by integration. Such a method can mitigate the need for validation data and can reduce bias in estimation of association by fixing sensitivity and specificity at particular values. With simulation studies and a real data example from Kaiser Permanente Washington (KPW), an integrated health care system in Washington State, we demonstrate the advantage of this proposed method over existing methods.

METHODS

We first compare the bias of the estimated association parameters obtained from PIE and 2 commonly used methods using simulated data. Then we evaluate the performance of the 3 methods on an EHR dataset with information about type 2 diabetes from KPW, where gold standard information (defined in the description of dataset) is available.

Development and evaluation of the PIE using simulated data

Simulation settings

To illustrate the idea in its simplest form, we consider a setting with only one risk factor. However, proposed methods apply to more complex settings that include multiple predictors. We wished to study the association between a continuous predictor, x (eg, number of cigarettes per day for one person), and a binary disease outcome, y (eg, type II diabetes), using EHR-derived data. Due to imperfect phenotyping, the identified diabetes status is subject to misclassification, ie, a surrogate variable, S_i, is observed rather than the true disease status, Y_i, where i is the index of the subject. We assume the true association between x_i and Y_i is described by a logistic regression model

$$\text{logit}\{\Pr(Y_i = 1)\} = \beta_0 + \beta_1 * x_i, \quad (1)$$

where $\text{logit}(p) = \log\{p/(1-p)\}$. In the nondifferential misclassification scenario, ie, where the misclassification rates of the surrogate are not modified by the exposure level, the relationship between x_i, and the surrogate variable, S_i, can be described as

$$\Pr(S_i = 1) = (1-\alpha_0) + (\alpha_0 + \alpha_1 - 1)\text{expit}(\beta_0 + \beta_1 * x_i), \quad (2)$$

where $\text{expit}(p) = \exp(p)/\{1+\exp(p)\}$, $\alpha_1 = \Pr(S_i = 1|Y_i = 1)$ and $\alpha_0 = \Pr(S_i = 0|Y_i = 0)$ are the sensitivity and specificity of the phenotyping algorithm, respectively.

We considered scenarios with disease prevalence ranging from 20% to 80% and 2 values of effect size, $\beta_1 = 1$ and 1.5, in model (1). The sensitivity and specificity of a phenotyping algorithm for the disease were either high (0.85 and 0.90, respectively) or low (0.65 and 0.80, respectively). The continuous predictor was generated from a normal distribution for 1000 individuals, $x_i \sim N(0,\ \sigma^2)$, where $\sigma^2 = 4$, $i = 1, \ldots 1000$. The true disease status of each subject was generated from a binomial distribution with success rate, $\Pr(Y_i = 1)$, calculated using model (1). The observed surrogate, S_i, was then generated using the assumed misclassification rates.

Algorithms

The association parameter, β_1, can be estimated using the following methods:

Method ignoring phenotyping errors (naive). A straightforward solution to estimating the odds ratio, β_1, is to ignore misclassification

and treat the surrogate S_i as the true disease status. This estimates the regression coefficient γ_1 in the logistic regression model

$$\text{logit}\{\Pr(S_i = 1)\} = \gamma_0 + \gamma_1 * x_i. \quad (3)$$

Although this method is simple and easy to implement, the estimated association $\hat{\gamma}_1$ is a biased estimate of the true association, and is toward null under nondifferential misclassification.[22]

ML method, unknown accuracy. A more rigorous procedure is to use the MLE, which treats misclassification rates as nuisance parameters jointly estimated with the association parameters. In the nondifferential misclassification scenario, the likelihood function is constructed as

$$L(\beta_0, \beta_1, \alpha_0, \alpha_1) = \prod_{i=1}^{n} p_i^{S_i}(1 - p_i)^{1-S_i}, \quad (4)$$

where $p_i = \Pr(S_i = 1) = (1 - \alpha_0) + (\alpha_0 + \alpha_1 - 1)\text{expit}(\beta_0 + \beta_1 x_i)$. The parameter of interest β_1 can be estimated by maximizing the likelihood $L(\beta_0, \beta_1, \alpha_0, \alpha_1)$. The advantage of this method is that the misclassified binary outcome is modeled using α_1 and α_0, and the MLE is guaranteed to be unbiased when the sample size is very large. However, the practical utility of this approach is limited by the need for extremely large sample sizes.[14,15] The performance of the MLE in moderate sample sizes is poor, because the shape of the likelihood $L(\beta_0, \beta_1, \alpha_0, \alpha_1)$ is usually very flat, leading to bias, as shown in Figure 1. Thus, this method is not commonly used in practice.

ML method, conditioned on accuracy (ML with fixed accuracy parameters). To reduce the bias caused by the undesirable performance of the MLE, one can fix the sensitivity and specificity at given values and maximize the resulting likelihood function. For example, by fixing $\alpha_0 = 0.90$ and $\alpha_1 = 0.85$, the new likelihood function becomes

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{S_i}(1 - p_i)^{1-S_i}, \quad (5)$$

where $p_i = 0.1 + 0.75\text{expit}(\beta_0 + \beta_1 x_i)$. The parameter of interest β_1 can be estimated by maximizing the likelihood $L(\beta_0, \beta_1)$. The disadvantage of this method is that correct specification of sensitivity and specificity requires a large validation sample, which is not cost-effective, and misspecification of these accuracy parameters will lead to biased estimation of β_1.

Prior knowledge guided integrated likelihood estimation method (PIE). IL is a novel tool developed recently to make valid inferences for parameters of interest in the presence of nuisance parameters.[21,23] It eliminates the nuisance parameters (here, sensitivity and specificity) by integrating with respect to a prior function, so that the resultant IL depends only on the parameters of interest (here, the regression coefficients) and the data. Unlike standard likelihood-based inference, where the nuisance parameters are maximized over their ranges, in the IL the nuisance parameters are "averaged" or "smoothed" over their ranges.[21,23] The resultant likelihood function, $L_I(\beta_0, \beta_1)$, can be used as a standard likelihood function for inference under certain conditions, and the estimate is obtained by maximizing the IL. We propose to use a PIE method to account for misclassification of phenotypes and to correct estimation bias in EHR-based association studies.

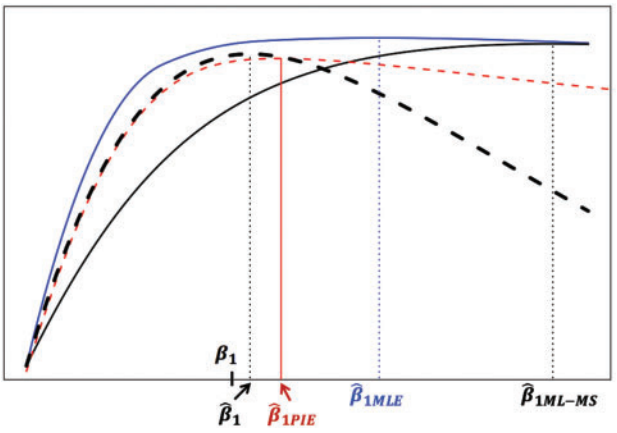

Figure 1. Comparison of likelihood function with unknown accuracy (blue solid line), likelihood function conditioned on misspecified accuracy (black solid line), likelihood function conditioned on known accuracy (black dashed line), and prior knowledge guided integrated likelihood function (red solid line). The true sensitivity and specificity are 90%.

Figure 1 illustrates the key idea of PIE and contrasts it with existing approaches. We generated true disease status based on model (1) and a surrogate with sensitivity and specificity of 90%. The true value of the association parameter is β_1. We compare the shape of the likelihood functions and the estimates of β_1 when different approaches are used. When the sensitivity and specificity are correctly specified, the estimate $\hat{\beta}_1$ is close to the true value β_1. When the sensitivity and specificity are incorrectly specified (both fixed at 100%), the estimate $\hat{\beta}_{1\text{ML}-\text{MS}}$ has large bias. When the sensitivity and specificity are unknown, the likelihood based on the joint function for the association and misclassification parameters reaches its plateau over a wide range. As a consequence, the ML estimate $\hat{\beta}_{1\text{MLE}}$ (in blue) is grossly biased. However, the proposed IL (in red) is much more quadratic than the likelihood (in blue), and the maximum IL estimate, $\hat{\beta}_{1\text{PIE}}$ (in red), is close to the true value.

Practically, to account for phenotyping errors and reduce information bias, the proposed PIE method can be conducted in 2 steps:

1. Construct prior distributions for sensitivity and specificity from a small validation study or a literature review on available evidence for accuracy of phenotyping algorithms, and
2. Incorporate the prior distribution into the likelihood using the PIE method to achieve bias reduction.

In practice, the exact sensitivity and specificity of a phenotyping algorithm are often unknown. However, a reasonable range or distribution of the sensitivity and specificity can be obtained by mining the existing literature or analyzing a small validation study.

The proposed IL is constructed as

$$L_I(\beta_0, \beta_1) = \int\int L(\beta_0, \beta_1, \alpha_0, \alpha_1)\pi(\alpha_0, \alpha_1)d\alpha_0 d\alpha_1,$$

where $\pi(\alpha_0, \alpha_1)$ is a prior distribution for sensitivity and specificity. We note that formulating the IL does not require accurate specification of the sensitivity and specificity at a particular value. A plausible range or distribution is adequate. This feature makes the proposed method feasible in practical settings and robust to misspecification of the sensitivity and specificity, which can minimize the cost of extensive chart review while reducing information bias.

Table 1. Five prior distributions used for the proposed PIE method

Prior names	Prior for sensitivity		Prior for specificity	
	Distribution	sd	Distribution	sd
PIE1	0.5+1/2*logitnormal(0.67,0.60)	0.07	0.5+1/2*logitnormal(0.73,0.80)	0.08
PIE1_sv	0.5+1/2*logitnormal(0.70,0.20)	0.02	0.5+1/2*logitnormal(0.80,0.23)	0.03
PIE2	0.5+1/2*logitnormal(0.50,0.60)	0.07	0.5+1/2*logitnormal(0.58,0.60)	0.07
PIE2_lv	0.5+1/2*logitnormal(0.50,1.20)	0.12	0.5+1/2*logitnormal(0.53,1.20)	0.12
PIE3	uniform(0.60,0.90)	0.09	uniform(0.65,0.95)	0.09

Experiments and evaluation

In this study, we compare the PIE method to the method ignoring phenotyping errors, referred to hereafter as the naïve method, and the ML method with misspecified sensitivity and specificity, referred to hereafter as the ML with misspecification, or the ML-MS method. The ML method with unknown phenotyping errors was not included in the comparison, as this method is not commonly used and is not considered practical. We also included the ML method with known accuracy, referred to hereafter as the gold standard method. In reality, such a situation is relatively rare.

We simulated 500 datasets and compared the bias of the estimated β_1 from the naïve method, the ML-MS method, the gold standard, and the proposed PIE method. For the PIE method, we evaluated the performance under 5 prior distributions as follows, also shown in Table 1 and Figure 2.

1. PIE1: transformed logit normal prior distributions with highest density around the true values of sensitivity and specificity.
2. PIE1_sv: transformed logit normal prior distributions with highest density around the true values of sensitivity and specificity, with small variances (sv).
3. PIE2: transformed logit normal prior distributions with highest density ~10% (on the scale of sensitivity and specificity) different from the true values of sensitivity and specificity.
4. PIE2_lv: transformed logit normal prior distributions with highest density ~10% (on the scale of sensitivity and specificity) different from the true values of sensitivity and specificity, with large variances (lv).
5. PIE3: uniformly distributed prior distribution with a range of 30%, centered ~10% (on the scale of sensitivity and specificity) different from the true values of sensitivity and specificity.

The first 2 priors mimic the situation where the phenotyping algorithm has been previously applied in similar settings and the performance of the algorithm is relatively well understood. The second 2 priors mimic the case where the phenotyping algorithm is less well characterized or its performance differs across datasets. In such situations, the highest density of the prior distribution obtained from previous studies deviates from the actual performance in the study population. The last prior mimics a situation in which investigators believe that the phenotyping error could be any value within a range with equal probability, a common situation in practice. For comparability of PIE and ML-MS, we set the sensitivity and specificity for the ML-MS method to values that are the same as the maximum of the third and fourth prior distributions (PIE2 and PIE2_lv). We calculated the mean and variance of estimation bias for each method as the mean and variance of the 500 estimates minus the true value of the association parameter.

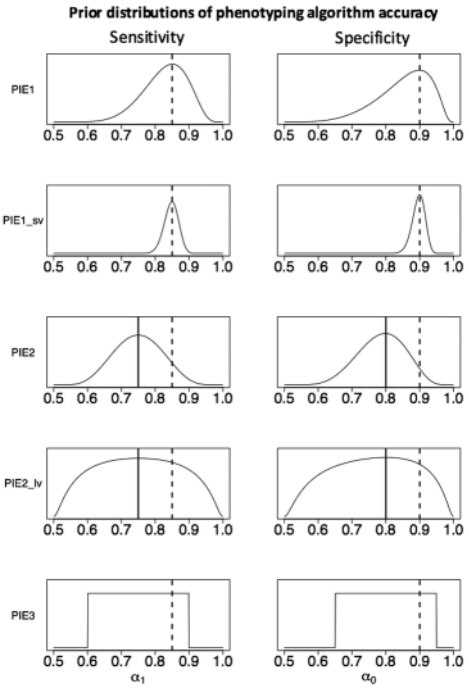

Figure 2. Illustration of the 5 types of prior distributions in PIE method: PIE1 (distributions peak at the true values of sensitivity and specificity); PIE1_sv (distributions peak at the true values of sensitivity and specificity, with small variance); PIE2 (distributions have peaks that differ from true values); PIE2_lv (distributions have peaks that differ from true values, with large variance); and PIE3 (uniform distributions not centered at the true values). Vertical dashed line marks the true value of sensitivity or specificity, and solid line marks the peak of the prior distribution.

Application of PIE to an EHR dataset including type 2 diabetes

Dataset

We applied the proposed method to a dataset derived from EHR data for a sample from KPW. Data were provided by the Adult Changes in Thought study, a longitudinal study of aging and dementia. Participants were dementia-free, at least 65 years old at the time of study enrollment, and randomly selected from the KPW membership. Study procedures have been previously described.[24] Our analysis was based on a deidentified subset consisting of 2022 participants who met the same inclusion criteria as a prior study of glucose and dementia.[16]

In the current analysis, "treated diabetes" was the phenotype of interest and the gold standard was defined as "two filled prescriptions for diabetes medications." Based on KPW pharmacy records, we

extracted this information for all 2022 participants. An imperfect surrogate measure for treated diabetes was created by dichotomizing the average glucose level in the prior 5 years, based on laboratory results for glucose and hemoglobin A1c, using a threshold of 140 mg/dL. We investigated the association between treated diabetes and predictors of interest, namely body mass index (BMI), treated hypertension, and race (white vs nonwhite). By comparing the surrogate and true diabetes measures, we estimated the true sensitivity and specificity of the surrogate to be 0.89 and 0.98, respectively.

Evaluation

We applied the naïve method, the ML method with true sensitivity and specificity (gold standard in the simulation section), the ML-MS method (accuracy 5% lower than the true sensitivity and specificity), and the PIE method to this dataset. We used uniform prior distributions with ranges from 0.80 to 0.99 for sensitivity and specificity in the PIE method, resembling the scenario where, based on prior studies, investigators hypothesize the misclassification rates of their phenotype to be at least 0.80. We compare the relative bias of the estimated effect sizes (log odds ratio) for BMI, hypertension, and race on type 2 diabetes using the 4 methods.

RESULTS

Evaluation of bias reduction through simulation studies

To illustrate the specification of the prior distributions in the PIE method, Figure 2 visualizes the priors where true sensitivity and specificity of a phenotyping algorithm are 85 and 90%, respectively.

Figure 3 presents comparison of the estimates of β_1 using box plots. As expected, the gold standard method yielded estimates with almost no bias and small variance. The estimates from the naïve method had small variance but large bias toward the null. The estimates of the ML-MS method had both large bias and large variation. In contrast, the bias of the proposed PIE method under all 3 prior distributions (PIE1, PIE2, and PIE3) was substantially smaller than that of the naïve and ML-MS methods. Under PIE2, when the peak of the prior distribution was about 10% lower than the truth, the proposed PIE method had smaller bias compared to the ML-MS method. This finding reveals the key advantage of the PIE method: even when the prior distributions of sensitivity and specificity do not peak at the true values, the PIE method can still reduce the bias by integrating over the possible values of sensitivity and specificity. Interestingly, PIE3 (with a uniform prior not centered at the truth) has much smaller bias than PIE2, and has comparable bias and only slightly larger variance than PIE1 (where the prior distribution is peaked at the truth). This suggests that (1) strong nonuniform priors not peaking at the truth can lead to some bias, and (2) strong nonuniform priors peaking at the truth sometimes cannot lead to much efficiency gain compared to a weak uniform prior. Such findings shed light on better strategies for specifying priors for PIE methods.

By comparing the results from the PIE methods with the naïve method, we can see a clear variance-bias trade-off. However, the bias of the naïve method persists in larger samples, while the variance of the PIE estimates becomes smaller with larger sample size.

Table 2 provides a more quantitative comparison of bias and variance among the methods under evaluation. Compared to the naive method, the percentage of relative bias reduction (absolute bias reduction divided by true association) of the PIE methods is between 38% and 65%, 20% and 65%, and 28% and 47%, when the prior distribution is peaked at the truth (PIE1), peaked at 10% away from the truth (PIE2), and uniformly distributed with center not at the truth (PIE3), respectively. The PIE method can reduce bias more when the true association is stronger, ie, $\beta_1 = 1.5$ compared to $\beta_1 = 1$, or when the actual sensitivity and specificity are lower, ie, $\alpha_0/\alpha_0/\alpha_1 = 80\%/65\%$ compared to $\alpha_1 = 90\%/85\%$. Compared to the ML-MS method, the percentage of relative bias reduction of the PIE methods is up to 78%. The standard deviations of the estimates of the PIE methods also increase when the true association is stronger (with difference up to 0.23) and the actual sensitivity and specificity are lower (with difference up to 0.41).

Figure 4 shows the relative impact of bias and variance of prior distributions on the performance of PIE estimates. We found that

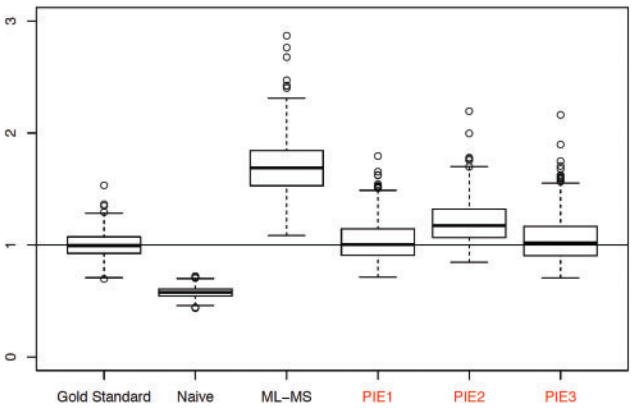

Figure 3. Box plots of estimates of β_1 using the ML method with correctly specified sensitivity and specificity (gold standard), the method ignoring misclassification (naïve), the ML method with misspecified sensitivity and specificity (ML-MS), and the prior knowledge guided integrated likelihood method with 3 priors (PIE1, PIE2, PIE3). Solid black segment in each box shows the median of the estimates.

Table 2. Comparison of methods for estimation of the association parameter, β_1, in term of bias and standard deviation

α_1, α_0	β_1	Bias						Standard deviation					
		GS	Naïve	ML-MS	PIE1	PIE2	PIE3	GS	Naïve	ML-MS	PIE1	PIE2	PIE3
0.85, 0.90	1	0.00	−0.42	0.70	0.04	0.21	0.06	0.10	0.09	0.24	0.17	0.20	0.20
	1.5	0.03	−0.78	1.24	0.07	0.24	0.08	0.17	0.08	0.47	0.28	0.28	0.30
0.65, 0.80	1	0.04	−0.70	−0.45	0.11	−0.23	−0.42	0.22	0.09	0.10	0.51	0.39	0.14
	1.5	0.08	−1.15	−0.79	0.17	−0.17	−0.68	0.44	0.08	0.10	0.69	0.62	0.22

Abbreviations: GS, or gold standard: ML method with true sensitivity and specificity; Naive: method ignoring misclassification; ML-MS: ML method with misspecified sensitivity and specificity; PIE1, PIE2, PIE3: PIE methods under 3 priors.

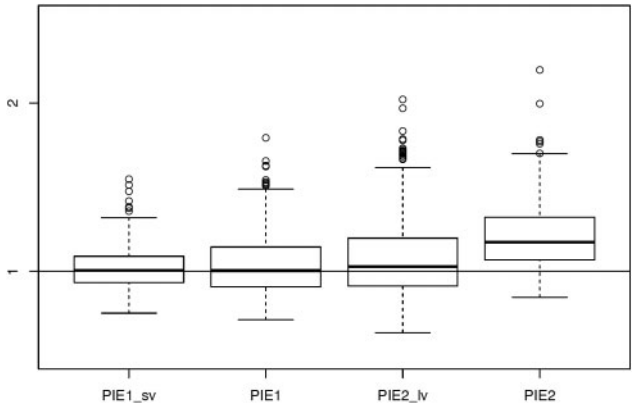

Figure 4. Box plots of estimates of β_1 using the prior knowledge guided integrated likelihood method with 4 priors (PIE1_sv, PIE1, PIE2_lv, PIE2). Solid black segment in each box shows the median of the estimates.

Table 3. Summary statistics of the variables of interest in the diabetes dataset from KPW

Variables of interest	$N = 2022$ (%)
Treated diabetes	
Yes	230 (11.4)
No	1792 (88.6)
Hypertension	
Yes	1403 (69.4)
No	619 (30.6)
Race	
White	1821 (90.1)
Nonwhite	201 (9.9)

when prior distributions were centered around the true value and had small variance (PIE1_sv), indicating that prior knowledge was accurate and informative, the PIE estimates had the best performance, with little bias and small variance. When the variance of the prior distribution increased (PIE1), the PIE estimates still had little bias but the variance increased. The worst scenario was when the prior distributions were centered away from the true value and had small variance (PIE2_lv), indicating that prior knowledge was inaccurately assumed to be informative, and the PIE estimates had large bias. When the variance of the prior distribution increased, the bias of the PIE estimates decreased but the variance increased, demonstrating a bias-variance trade-off.

Validation using the diabetes dataset from KPW

We applied the proposed method to the diabetes data from KPW. Table 3 summarizes the variables of interest in this dataset. Table 4 presents the estimated effects (on log odds ratio scale) of the risk factors for treated diabetes using different methods. Relative to the estimates of the gold standard method (using true sensitivity and specificity calculated based on true treated diabetes status), the estimated effects using the naïve method were biased toward the null for all 3 predictors. The relative biases were −23%, −11%, and −16% for hypertension, BMI, and race, respectively. The ML-MS method also had large bias, with inflated estimated effects relative to the estimates based on true treated diabetes status. PIE greatly reduced the bias, and the estimated effect sizes were very close to the gold standard method with relative bias <10%.

DISCUSSION

In this paper, we proposed PIE as a method to correct bias in association estimates due to information bias in EHR-derived data. The results of both simulation studies and real data analysis show that the proposed PIE method effectively reduced bias in estimation of associations by incorporating prior information on performance of phenotyping algorithms. The proposed method outperformed 2 existing methods that are commonly used in EHR-related studies and was comparable to the gold standard method. A unique strength of the proposed method is that it does not require specification of fixed values for sensitivity and specificity, thus is more robust to model misspecification compared to existing methods.

An important implication of the PIE method is that bias reduction without validation data is possible under practical scenarios for EHR-based studies. More precisely, when validation data are not available,[5,25–28] prior information on sensitivity and specificity can be obtained by mining the existing literature to extract previously estimated misclassification rates. For diseases that have been well studied and for which sensitivity and specificity of the phenotyping algorithms have been reported in various datasets (eg,[29]), the prior distribution of the sensitivity and specificity can be built using the empirical distribution of the sensitivity/specificity obtained from text-mining existing literature. In other scenarios where the condition is less studied or algorithms are newly developed such that prior information is limited in the literature, a uniform prior distribution with a reasonable range of values for sensitivity and specificity can be used. In both situations, the proposed method can substantially reduce the bias of estimated associations compared to the naïve method and the ML-MS method, as we demonstrated in simulation studies and a real case study.

In practice, when a small validation dataset is available,[1,28,30–36] a common strategy is to jointly model the validation data and the nonvalidated data, and base inferences on ML estimation. In future studies, it will be of interest to compare the performance of the ML estimation method with the PIE method, which incorporates information on phenotyping accuracy as an informative prior.

Further work is needed to fully develop and evaluate the PIE method. For example, the confidence sets for the PIE estimates can be obtained by reversing the IL ratio test.[19,21] Practically, such sets can also be obtained by resampling methods for computational efficiency. In addition, maximizing the IL function in PIE can be computationally expensive due to the double integration in the likelihood function when the dimension of predictors is relatively high. Numerical optimization approaches, eg, coordinate descent,[37–39] need to be developed to improve computational efficiency. Furthermore, we have not evaluated a full Bayesian approach in our methods comparison. It would be of interest to develop a full Bayesian method and compare it with the proposed PIE method. Finally, the current investigation has been limited to the case where misclassification is nondifferential. In practice, misclassification rates may depend on exposure status. The PIE method needs to be further developed to account for such challenges.

In this paper, we have focused on correction of bias due to misclassification of binary outcomes in EHR-derived data. Similar ideas can be adapted to misclassification of survival outcomes and measurement errors in exposure variables. These extensions are currently under investigation and will be reported in the future. We believe the proposed approach is an important contribution to bias reduction in EHR-based association studies.

Table 4. Estimated effect sizes (in log odds ratio scale) of the risk factors for diabetes using different methods

	Hypertension		BMI		Race	
	Point estimate	Relative bias (%)	Point estimate	Relative bias (%)	Point estimate	Relative bias (%)
Gold standard	0.53	0	0.09	0	0.57	0
Naïve	0.41	−23	0.08	−11	0.48	−16
ML-MS	0.68	28	0.10	11	0.65	14
PIE	0.48	−9	0.09	0	0.56	2

CONCLUSION

In this study, we proposed a maximum IL estimation method, the PIE method, to reduce estimation bias by incorporating prior knowledge of phenotyping errors. Our evaluation using simulated datasets and data from KPW demonstrated that the proposed PIE method can effectively reduce bias compared to methods that are commonly used in current EHR-based studies.

FUNDING

Research reported in this publication was supported in part by the National Institutes of Health (NIH) under award numbers R01AI130460, R01GM103859, U24CA194215, U01AG006781, ES013508, DK112217, and TR001878, the Patient-Centered Outcomes Research Institute (PCORI) under award number ME-1511-32666, and a Commonwealth Universal Research Enhancement (CURE) Program grant from the Pennsylvania Department of Health. All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of NIH, PCORI, the PCORI Board of Governors, or the PCORI Methodology Committee.

CONTRIBUTORS

JH, RD, RH, YW, JM, HX, and YC designed methods and experiments; RH provided the dataset from Kaiser Permanente Washington for data analysis; RH, YW, and YC guided the dataset generation for the simulation study; JH and RD generated the simulation datasets, conducted simulation experiments, and conducted data analysis of the EHR data from Kaiser Permanente Washington; RH, JM, HX, and YC interpreted the results and provided instructive comments; JH, RD, RH, and YC drafted the main manuscript. All authors have approved the manuscript.

COMPETING INTERESTS

The authors have no competing interests to declare.

ACKNOWLEDGMENT

We wish to thank Dr Paul Crane and Dr Eric Larson, PIs of the Adult Changes in Thought study, for providing data for analysis.

REFERENCES

1. Denny JC, Crawford DC, Ritchie MD, *et al*. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Human Genet*. 2011;89(4):529–42.
2. Denny JC, Ritchie MD, Crawford DC, *et al*. Identification of genomic predictors of atrioventricular conduction using electronic medical records as a tool for genome science. *Circulation*. 2010;122(20):2016–21.
3. Kho AN, Pacheco JA, Peissig PL, *et al*. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Trans Med*. 2011;3(79):79re1.
4. Lemke AA, Wu JT, Waudby C, *et al*. Community engagement in biobanking: experiences from the eMERGE Network. *Genomics, Soc Policy*. 2010;6(3):1–18.
5. Ritchie MD, Denny JC, Crawford DC, *et al*. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Human Genet*. 2010;86(4):560–72.
6. Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Stat Med*. 2001;20(1):139–60.
7. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323–37.
8. Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: what data are observed and why? *eGEMs*. 2016;4(1):1203.
9. Wei W-Q, Teixeira PL, Mo H, *et al*. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc*. 2016;23(e1):e20–27.
10. Denny JC, Ritchie MD, Basford MA, *et al*. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26(9):1205–10.
11. Rasmussen LV, Kiefer RC, Mo H, *et al*. A modular architecture for electronic health record–driven phenotyping. *AMIA Summits Transl Sci Proc*. 2015;2015:147.
12. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *Am J Epidemiol*. 1997;146(2):195–203.
13. Duan R, Cao M, Wu Y, *et al*. An empirical study for impacts of measurement errors on ehr based association studies. *AMIA Annu Symp Proc*. 2017;2016:1764–73.
14. Carroll RJ, Ruppert D, Stefanski LA, *et al*. *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, FL: CRC Press; 2006.
15. Copas JB. Binary regression models for contaminated data. *J Royal Stats Soc. Series B (Methodological)*. 1988;50:225–65.
16. Crane PK, Walker R, Hubbard RA, *et al*. Glucose levels and risk of dementia. *New Engl J Med*. 2013;369(6):540–48.
17. Luan X, Pan W, Gerberich SG, *et al*. Does it always help to adjust for misclassification of a binary outcome in logistic regression? *Stats Med*. 2005;24(14):2221–34.
18. Wei W-Q, Leibson CL, Ransom JE, *et al*. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc* 2012;19(2):219–24.
19. Khurshid S, Keaney J, Ellinor PT, *et al*. A simple and portable algorithm for identifying atrial fibrillation in the electronic medical record. *Am J Cardiol*. 2016;117(2):221–25.
20. Meier AS, Richardson BA, Hughes JP. Discrete proportional hazards models for mismeasured outcomes. *Biometrics*. 2003;59(4):947–54.
21. Severini TA. Integrated likelihood functions for non-Bayesian inference. *Biometrika*. 2007;94(3):529–42.

22. Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*. 1999;86(4):843–55.
23. Berger JO, Liseo B, Wolpert RL. Integrated likelihood methods for eliminating nuisance parameters. *Stats Sci*. 1999;14(1):1–28.
24. Kukull WA, Higdon R, Bowen JD, *et al*. Dementia and Alzheimer disease incidence: a prospective cohort study. *Arch Neurol*. 2002;59(11): 1737–46.
25. Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *BMJ*. 2009;338:b81.
26. Kohane IS, McMurry A, Weber G, *et al*. The co-morbidity burden of children and young adults with autism spectrum disorders. *PLoS One*. 2012;7(4):e33224.
27. Klompas M, Haney G, Church D, *et al*. Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance. *PLoS One*. 2008;3(7):e2626.
28. Navaneethan SD, Jolly SE, Schold JD, *et al*. Development and validation of an electronic health record–based chronic kidney disease registry. *Clin J Am Soc Nephrol*. 2011;6(1):40–9.
29. Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc*. 2012;19(e1):e162–9.
30. Liao KP, Cai T, Gainer V, *et al*. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res*. 2010;62(8):1120–27.
31. Desai JR, Wu P, Nichols GA, *et al*. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. *Med Care*. 2012;50:S30.
32. Parsons A, McCullough C, Wang J, *et al*. Validity of electronic health record–derived quality measurement for performance monitoring. *J Am Med Inform Assoc*. 2012;19(4):604–09.
33. Benin AL, Fenick A, Herrin J, *et al*. How good are the data? Feasible approach to validation of metrics of quality derived from an outpatient electronic health record. *Am J Med Qual*. 2011;26:441–51.
34. Newton KM, Peissig PL, Kho AN, *et al*. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE Network. *J Am Med Inform Assoc*. 2013;20(e1):e147–54.
35. Castro VM, Minnier J, Murphy SN, *et al*. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *Am J Psychiatry*. 2015;172(4):363–72.
36. Kho AN, Hayes MG, Rasmussen-Torvik L, *et al*. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc*. 2012;19(2):212–8.
37. Hildreth C. A quadratic programming procedure. *Naval Res Logistics*. 1957;4(1):79–85.
38. Warga J. Minimizing certain convex functions. *J Soc Indust Appl Math*. 1963;11(3):588–93.
39. Ortega J, Rheinboldt W. Iterative Solution of Nonlinear Equations in Several Variables. *Vol. 30*. Philadelphia: SIAM; 1970.